Student Models for Prior Knowledge Estimation

Juraj Nižnan
Masaryk University Brno
niznan@mail.muni.cz

Radek Pelánek
Masaryk University Brno
xpelanek@mail.muni.cz

Jiří Řihák
Masaryk University Brno
thran@mail.muni.cz

ABSTRACT

Intelligent behavior of adaptive educational systems is based on student models. Most research in student modeling focuses on student learning (acquisition of skills). We focus on prior knowledge, which gets much less attention in modeling and yet can be highly varied and have important consequences for the use of educational systems. We describe several models for prior knowledge estimation – the Elo rating system, its Bayesian extension, a hierarchical model, and a networked model (multivariate Elo). We evaluate their performance on data from application for learning geography, which is a typical case with highly varied prior knowledge. The result show that the basic Elo rating system provides good prediction accuracy. More complex models do improve predictions, but only slightly and their main purpose is in additional information about students and a domain.

1. INTRODUCTION

Computerized adaptive practice [14, 22] aims at providing students with practice in an adaptive way according to their skill, i.e., to provide students with tasks that are most useful to them. In this work we focus on the development of adaptive systems for learning of facts, particularly on modeling of prior knowledge of facts.

In student modeling [6] most attention is usually paid to modeling student learning (using models like Bayesian Knowledge Tracing [4] or Performance Factors Analysis [24]). Modeling of prior knowledge was also studied in prior work [22, 23], but it gets relatively little attention. It is, however, very important, particularly in areas where students are expected to have nontrivial and highly varying prior knowledge, e.g., in domains like geography, biology, human anatomy, or foreign language vocabulary. As a specific case study we use application for learning geography, which we developed in previous work [22]. The estimate of prior knowledge is used in models of current knowledge (learning), i.e., it has important impact on the ability of the practice system to ask suitable questions.

We consider several approaches to modeling prior knowledge and explore their trade-offs. The basic approach (described in previous work [22]) is based on a simplifying assumption of homogeneity among students and items. The model uses a global skill for students and a difficulty parameter for items; the prior knowledge of a student for a particular item is simply the difference between skill and difficulty. The model is basically the Rasch model, where the parameter fitting is done using a variant of the Elo rating system [9, 25] in order to be applicable in an online system.

The first extension is to capture the uncertainty in parameter estimates (student skill, item difficulty) by using Bayesian modeling. We propose and evaluate a particle based method for parameter estimation of the model. This approach is further extended to include multiplicative factors (as in collaborative filtering [15]) which allows to better model the heterogeneity among students and items.

The second extension is the hierarchical model which tries to capture more nuances of the domain by dividing items into disjoint subsets called concepts (or knowledge components). The model then computes student skill for each of these concepts. Since these concept skills are related, they are still connected by a global skill. With this model we have to choose an appropriate granularity of used concepts and find an assignment of items to these concepts. We use both manually determined concepts (e.g., "continents" in the case of geography) and concepts learned automatically from the data [19].

The third extension is a networked model, which bypasses the choice of concepts by modeling relations directly on the level of items. This model can be seen as a variation on previously proposed multivariate Elo system [7]. For each item we compute the most similar items (based on students' answers), e.g., in the geography application, knowledge of Northern European countries is correlated. Prior knowledge of a student for a particular item is in this model estimated based on previous answers to similar items (still using the global skill to some degree).

Extended models are more detailed than the basic model and can potentially capture student knowledge more faithfully. They, however, contain more parameters and the parameter estimation is more susceptible to the noise in data. We compare the described models and analyze their performance on a large data set from application for learning geography [22].

The results show that the studied extensions do bring an improvement in predictive accuracy, but the basic Elo system is surprisingly good. The main point of extension is thus in their additional parameters, which bring an insight into the studied domain. We provide several specific examples of such insight.

2. MODELS

Although our focus is on modeling knowledge of facts, in the description of models we use the common general terminology used in student modeling, particularly the notions of *items* and *skills*. In the context of geography application (used for evaluation) items correspond to locations and names of places and skill corresponds to knowledge of these facts.

Our aim is to estimate the probability that a student s knows an item i based on previous answers of students s to questions about different items and previous answers of other students to questions about item i. As a simplification we use only the first answer about each item for each student.

In all models we use the logistic function $\sigma(x) = \frac{1}{1+e^{-x}}$ as a link between a skill and a probability that a student answers correctly. In the case of multiple-choice questions the probability can be modeled by a shifted logistic function $\sigma(x,k) = 1/k + (1-1/k)\frac{1}{1+e^{-x}}$, where k is the number of options. We restrict our attention to online models (models that are updated after each answer). Such models can adapt to user behavior quickly and therefore are very useful in adaptive practice systems.

2.1 Basic Model

The basic model (described in previous work [22] and currently used in the online application) uses a key assumption that both students and studied facts are homogeneous. It assumes that students' prior knowledge in the domain can be modeled by a one-dimensional parameter.

We model the prior knowledge by the Rasch model, i.e., we have a student parameter θ_s corresponding to the global knowledge of a student s of a domain and an item parameter d_i corresponding to the difficulty of an item i. The probability that the student answers correctly is estimated using a logistic function of a difference between the global skill and the difficulty: $P(correct|\theta_s, d_i) = \sigma(\theta_s - d_i)$.

A common approach to the parameter estimation for the Rasch model is joint maximum likelihood estimation (JMLE). This is an iterative approach that is slow for large data, particularly it is not suitable for an online application, where we need to adjust estimates of parameters continuously.

In previous work [22, 25] we have shown that the parameter estimation can be done effectively using a variant of the Elo rating system [9]. The Elo rating system was originally devised for chess rating, but we can use it in student modeling by interpreting a student's answer on an item as a "match" between the student and the item. The skill and difficulty estimates are updated as follows:

$$\begin{aligned} \theta_s &:= \theta_s + K \cdot (correct - P(correct|\theta_s, d_i)) \\ d_i &:= d_i + K \cdot (P(correct|\theta_s, d_i) - correct) \end{aligned}$$

where *correct* denotes whether the question was answered correctly and K is a constant specifying sensitivity of the estimate to the last attempt. An intuitive improvement, which is used in most Elo extensions, is to use an "uncertainty function" instead of a constant K – the update should get smaller as we have more data about a student or an item. We use an uncertainty function $U(n) = \alpha/(1+\beta n)$, where n is the number of previous updates to the estimated parameters and α, β are meta-parameters.

2.2 Bayesian Model

In the basic model the uncertainty is modeled as a simple function of number of attempts. Such an approach is a simplification since some answers are more informative than others and thus the effect of answers on reduction of uncertainty should be differentiated. This can be done by using a Bayesian modeling approach. For this model we treat θ_s, d_i and *correct* as random variables. We can use Bayes' theorem for updating our beliefs about skills and difficulties:

$$P(\theta_s, d_i|correct) \propto P(correct|\theta_s, d_i) \cdot P(\theta_s, d_i)$$

We assume that the difficulty of an item is independent of a skill of a student and thus $P(\theta_s, d_i) = P(\theta_s) \cdot P(d_i)$. The updated beliefs can be expressed as marginals of the conditional distribution, for example:

$$P(\theta_s|correct) \propto P(\theta_s) \cdot \int_{-\infty}^{\infty} P(correct|\theta_s, d_i = y) \cdot P(d_i = y) \mathrm{d}y$$

In the context of rating systems for games, the basic Elo system has been extended in this direction, particularly in the Glicko system [11]. It models prior skill by a normal distribution and uses numerical approximation to represent the posterior by a normal distribution and to perform the update of the mean and standard deviation of the skill distribution using a closed form expressions. Another Bayesian extension is TrueSkill [12], which further extends the system to allow team competitions.

This approach is, however, difficult to modify for new situations, e.g., in our case we want to use the shifted logistic function (for modeling answers to multiple-choice questions), which significantly complicates derivation of equations for numerical approximation. Therefore, we use a more flexible particle based method to represent the skill distribution. The skill is represented by a skill vector $\boldsymbol{\theta_s}$, which gives the values of skill particles, and probability vector $\boldsymbol{p_s}$, which gives the probabilities of the skill particles (sums to 1). The item difficulty is represented analogically by a difficulty vector $\boldsymbol{d_i}$ and a probability vector $\boldsymbol{p_i}$. In the following text the notation $\boldsymbol{p_{s_k}}$ stands for the k-th element of the vector $\boldsymbol{p_s}$.

The skill and difficulty vectors are initialized to contain values that are spread evenly in a specific interval around zero. The probability vectors are initialized to proportionally reflect the probabilities of the particles in the selected prior distribution. During updates, only the probability vectors change, the vectors that contain the values of the particles stay fixed. Particles are updated as follows:

$$\begin{aligned} \boldsymbol{p_{s_k}} &:= \boldsymbol{p_{s_k}} \cdot \sum_{l=1}^{n} P(correct|\theta_s = \boldsymbol{\theta_{s_k}}, d_i = \boldsymbol{d_{i_l}}) \cdot \boldsymbol{p_{i_l}} \\ \boldsymbol{p_{i_l}} &:= \boldsymbol{p_{i_l}} \cdot \sum_{k=1}^{n} P(correct|\theta_s = \boldsymbol{\theta_{s_k}}, d_i = \boldsymbol{d_{i_l}}) \cdot \boldsymbol{p_{s_k}} \end{aligned}$$

After the update, we must normalize the probability vectors so that they sum to one. A reasonable simplification that avoids summing over the particle values is:

$$\begin{aligned} \boldsymbol{p_{s}}_k &:= \boldsymbol{p_{s}}_k \cdot P(correct|\theta_s = \boldsymbol{\theta_{s}}_k, d_i = E[\boldsymbol{d_i}]) \\ \boldsymbol{p_{i}}_l &:= \boldsymbol{p_{i}}_l \cdot P(correct|\theta_s = E[\boldsymbol{\theta_s}], d_i = \boldsymbol{d_{i}}_l) \end{aligned}$$

where $E[\boldsymbol{d_i}]$ ($E[\boldsymbol{\theta_s}]$) is the expected difficulty (skill) particle value (i.e. $E[\boldsymbol{d_i}] = \boldsymbol{d_i}^T \cdot \boldsymbol{p_i}$). By setting the number of particles we can trade off between precision on one hand and speed and memory requirements on the other hand.

Using the described particle model in a real-world application would require storing the probabilities for all the particles in a database. If we assume that our beliefs stay normal-like even after many observations then we can approximate each of the posteriors by a normal distribution. This approach is called assumed-density filtering [17]. Consequently, each posterior can be represented by just two numbers, the mean and the standard deviation. In this simplified model, each update requires the generation of new particles. We generate the particles in the interval $(\mu - 6\sigma, \mu + 6\sigma)$. Otherwise, the update stays the same as before. After the update is performed, the mean and the standard deviation are estimated in a standard way: $\mu_{\theta_s} := \boldsymbol{\theta_s}^T \cdot \boldsymbol{p_s}, \sigma_{\theta_s} := \|\boldsymbol{\theta_s} - \mu_{\theta_s}\|_2$.

The model can be extended to include multiplicative factors for items (q_i) and students (r_s), similarly to the Q-matrix method [1] or collaborative filtering [15]. Let k be the number of factors, then x passed in to the likelihood function $\sigma(x)$ has the form: $x = \theta_s - d_i + \sum_{j=1}^{k} q_{i,j} \cdot r_{s,j}$. The updates are similar, we only need to track more variables.

2.3 Hierarchical Model

In the next model, which we call 'hierarchical', we try to capture the domain in more detail by relaxing the assumption of homogeneity. Items are divided into disjoint sets – usually called 'concepts' or 'knowledge components' (e.g., states into continents). In addition to the global skill θ_s the model now uses also the concept skill θ_{sc}. We use an extension of the Elo system to estimate the model parameters. Predictions are done in the same way as in the basic Elo system, we just correct the global skill by the concept skill: $P(correct|\theta_s, \theta_{sc}, d_i) = \sigma((\theta_s + \theta_{sc}) - d_i)$. The update of parameters is also analogical (U is the uncertainty function and γ is a meta-parameter specifying sensitivity of the model to concepts):

$$\begin{aligned} \theta_s &:= \theta_s + U(n_s) \cdot (correct - P(correct|\theta_s, \theta_{sc}, d_i)) \\ \theta_{sc} &:= \theta_{sc} + \gamma \cdot U(n_{sc}) \cdot (correct - P(correct|\theta_s, \theta_{sc}, d_i)) \\ d_i &:= d_i + U(n_i) \cdot (P(correct|\theta_s, \theta_{sc}, d_i) - correct) \end{aligned}$$

This proposed model is related to several student modeling approaches. It can be viewed as a simplified Bayesian network model [3, 13, 16]. In a proper Bayesian network model we would model skills by a probability distribution and update the estimates using Bayes rule; equations in our model correspond to a simplification of this computation using only point skill estimates. Bayesian network model can also model more complex relationships (e.g., prerequisites), which are not necessary for our case (fact learning). Other related modeling approaches are the Q-matrix method [1], which focuses on modeling mapping between skills and items (mainly using $N : M$ relations), and models based on knowledge space theory [8]. Both these approaches are more complex than the proposed model. Our aim here is to evaluate whether even a simple concept based model is sensible for modeling factual knowledge.

The advantage of the hierarchical model is that user skill is represented in more detail and the model is thus less sensitive to the assumption of homogeneity among students. However, to use the hierarchical model, we need to determine concepts (mapping of items into groups). This can be done in several ways. Concepts may be specified manually by a domain expert. In the case of geography learning application some groupings are natural (continents, cities). In other cases the construction of concepts is more difficult, e.g., in the case of foreign language vocabulary it is not clear how to determine coherent groups of words. It is also possible to create concepts automatically or to refine expert provided concepts with the use of machine learning techniques [5, 19].

To determine concepts automatically it is possible use classical clustering methods. For our experiments we used spectral clustering method [27] with similarity of items i, j defined as a Spearman's correlation coefficient c_{ij} of correctness of answers (represented as 0 or 1) of shared students s (those who answered both items). To take into account the use of multiple-choice questions we decrease the binary representation of a response r by guess factor to $r - 1/k$ (k is the number of options). Disadvantages of the automatic concept construction are unknown number of concept, which is a next parameter to fit, and the fact that found concepts are difficult to interpret.

It is also possible to combine the manual and the automatic construction of concepts [19]. With this approach the manually constructed concepts are used as item labels. Items with these labels are used as a training set of a supervised learning method (we used logistic regression with regularization). For the item i, the vector of correlation with all items c_{ij} is used as vector of features. Errors of the used classification method are interpreted as "corrected" labels; see [19, 20] for more details.

2.4 Networked Model

The hierarchical model enforces hard division of items into groups. With the next model we bypass this division by modeling directly relations among individual items, i.e., we treat items as a network (and hence the name 'networked model'). For each item we have a local skill θ_{si}. For each pair of items we compute the degree to which they are correlated c_{ij}. This is done from training data or – in the real system – once a certain number of answers is collected. After the answer to the item i all skill estimates for all other items j are updated based on c_{ij}. The model still uses the global skill θ_s and makes the final prediction based on the weighted combination of global and local skill: $P(correct|\theta_s, \theta_{si}) = \sigma(w_1\theta_s + w_2\theta_{si} - d_i)$. Parameters are updated as follows:

$$\begin{aligned} \theta_s &:= \theta_s + U(n_s) \cdot (correct - P(correct|\theta_s, \theta_{si})) \\ \theta_{sj} &:= \theta_{sj} + c_{ij} \cdot U(n_s) \cdot (correct - P(correct|\theta_s, \theta_{si})) \\ &\quad \text{for all items } j \\ d_i &:= d_i + U(n_i) \cdot (P(correct|\theta_s, \theta_{si}) - correct) \end{aligned}$$

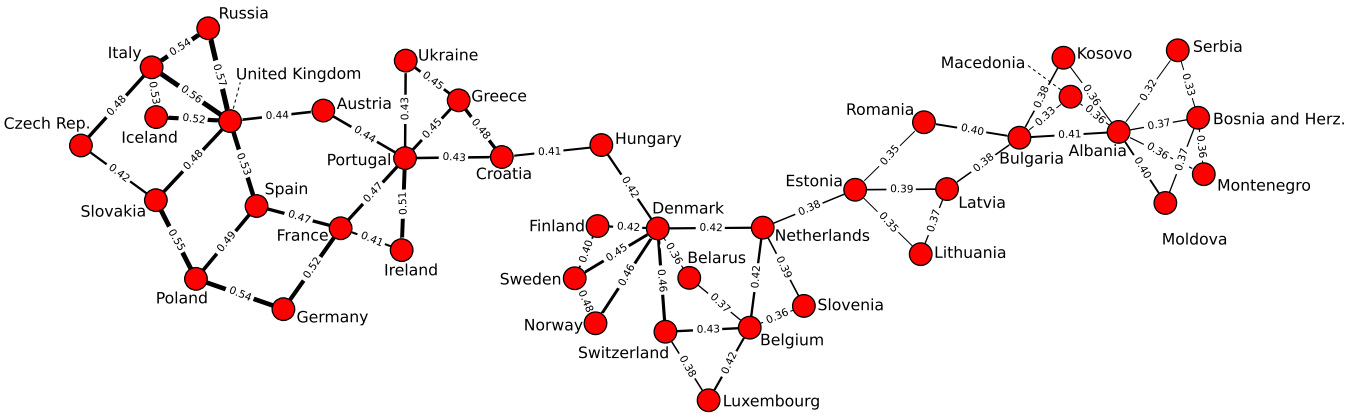

Figure 1: Illustration of the networked model on European countries. Only the most important edges for each country are shown.

This model is closely related to multivariate Elo which was previously proposed in the context of adaptive psychometric experiments [7].

For illustration of the model, Figure 1 shows selection of the most important correlations for European countries. Note that this automatically generated figure contains some natural clusters as Balkan countries (right), Scandinavian countries (middle), and well-known[1] countries (left).

3. EVALUATION

We provide evaluation of the above described models over data from an adaptive application for learning facts.

3.1 The Used System and Data

For the analysis we use data from an online adaptive system `slepemapy.cz` for practice of geography facts (e.g., names and location of countries, cities, mountains). The system estimates student knowledge and based on this estimate it adaptively selects questions of suitable difficulty [22]. The system uses a target success rate (e.g., 75 %) and adaptively selects questions in such a way that the students' achieved performance is close to this target [21]. The system uses open questions ("Where is France?") and multiple-choice questions ("What is the name of the highlighted country?") with 2 to 6 options. Students answer questions with the use of an interactive 'outline map'. Students can also access a visualization of their knowledge using an open learner model.

Our aim is to model prior knowledge (not learning during the use of the system), so we selected only the first answers of students to every item. The used data set contains more than 1.8 million answers of 43 thousand students. The system was originally available only in Czech, currently it is available in Czech, English, and Spanish, but students are still mostly from Czech republic ($> 85\%$) and Slovakia ($> 10\%$). The data set was split into train set (30%) and test set (70%) in a student-stratified manner. As a primary metric for model comparison and parameter fitting we use root mean square error (RMSE), since the application works with absolute values of predictions [22] (see [26] for more details on choice of a metric).

[1]By students using our system.

3.2 Model Parameters

The train set was used for finding the values of the meta-parameters of individual models. Grid search was used to search the best parameters of the uncertainty function $U(n)$. Left part of Figure 2 shows RMSE of the basic Elo model on training data for various choices of α and β. We chose $\alpha = 1$ and $\beta = 0.06$ and we used these values also for derived models which use the uncertainty function.

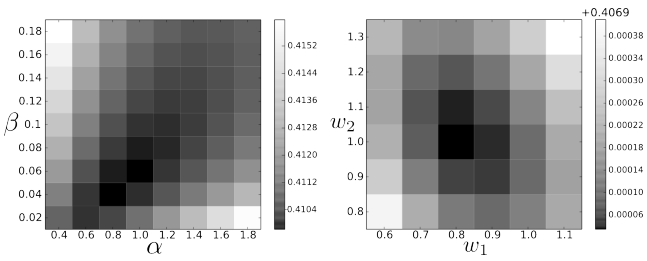

Figure 2: Grid searches for the best uncertainty function parameters α, β (left) and the best parameters w_1, w_2 of the networked model (right). As can be seen from different scales, models are more sensitive to α and β parameters.

Grid search (Figure 2 right) was used also to find the best parameters $w_1 = 0.8, w_2 = 1$ of the networked model. The train set was also used for computation of correlations. To avoid spurious high correlations of two items i, j as consequence of lack of common students we set all $c_{ij} = 0$ for those pairs i, j with less than 200 common students. Correlations computed by this method show stability with respect to selection of train set. For two different randomly selected train sets correlation values correlate well (> 0.95). As Figure 1 shows, the resulting correlations are interpretable.

For the particle-based Bayesian model we can tune the performance by setting the number of particles it uses for estimating each distribution. We found out that increasing the number of particles beyond 100 does not increase performance. For the simplified version, only 10 particles are sufficient. This is probably due to the way the algorithm uses the particles (they are discarded after each step).

Table 1: Comparison of models on the test set.

Model	RMSE	LL	AUC
Elo ($\alpha = 1, \beta = 0.06$)	0.4076	−643179	0.7479
Bayesian model	0.4080	−644362	0.7466
Bayesian model (3 skills)	0.4056	−637576	0.7533
Hierarchical model	0.4053	−636630	0.7552
Networked model	0.4053	−636407	0.7552

3.3 Accuracy of Predictions

All the reported models work online. Training of models (parameters θ_s, d_i) continues on the test set but only predictions on this set are used to evaluate models.

Table 1 shows results of model comparison with respect model performance metrics. In addition to RMSE we also report log-likelihood (LL) and area under the ROC curve (AUC); the main result are not dependent on the choice of metric. In fact, predictions for individual answers are highly correlated. For example for the basic Elo model and hierarchical model most of the predictions (95%) differ by less than 0.1.

The hierarchical model reported in Table 1 uses manually determined concepts based on both location (e.g., continent) and type of place (e.g., country). Both the hierarchical model and the networked model bring an improvement over the basic Elo model. The improvement is statistically significant (as determined by a t-test over results of repeated cross-validation), but it is rather small. Curiously, the Particle Bayes model is slightly worse than the simple Elo system, i.e., the more involved modeling of uncertainty does not improve predictions. The performance improves only when we use the multiple skill extension. We hypothesize that the improvement of the hierarchical (resp. multiple skill) extensions model be more significant for less homogeneous populations of students. Each skill could then be used to represent a different prior knowledge group.

RMSE is closely related to Brier score [26], which provides decomposition [18] to uncertainty (measures the inherent uncertainty in the observed data), reliability (measures how close the predictions are to the true probabilities) and resolution (measures how diverse the predictions are).

This decomposition can be also illustrated graphically. Figure 3 shows comparison of the basic Elo model and the hierarchical model. Both calibration lines (which are near the optimal one) reflect very good reliability. On the other hand, histograms reflect the fact that the hierarchical model gives more divergent predictions and thus has better resolution.

3.4 Using Models for Insight

In student modeling we are interested not just in predictions, but also in getting insight into characteristics of the domain or student learning. The advantage of more complex models may lie in additional parameters, which bring or improve such insight.

Figure 5 gives comparison of item difficulty for Elo model

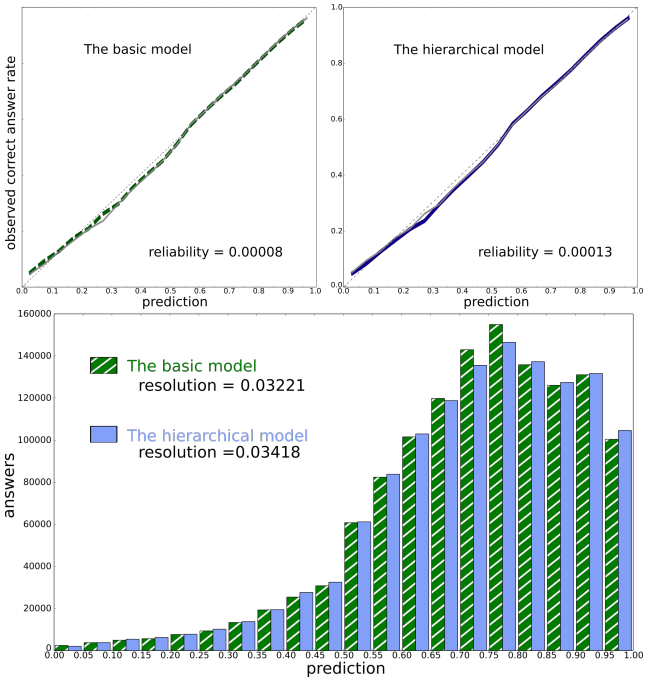

Figure 3: Illustration of the Brier score decomposition for the basic model and the hierarchical model. Top: reliability (calibration curves). Bottom: resolution (histograms of predicted values).

and Particle Bayes. As we can see, the estimated values of the difficulties are quite similar. The main difference between these models is in estimates of uncertainty. The uncertainty function used in Elo converges to zero faster and its shape is the same for all items. In Particle Bayes, the uncertainty is represented by the standard deviation of the normal distribution. This uncertainty can decrease differently for each item, depending on the amount of surprising evidence the algorithm receives, as is shown in Figure 4. The better grasp of uncertainty can be useful for visualization in an open learner model [2, 10].

Other extensions (networked, hierarchical, Bayesian with multiple skills) bring insight into the domain thanks to the analysis of relations between items, e.g., by identifying most

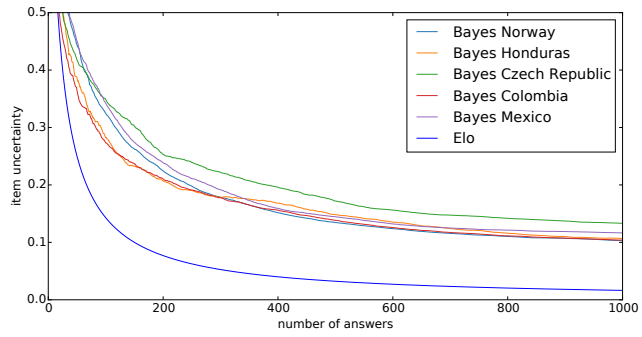

Figure 4: Evolution of uncertainties in the Bayes model and Elo.

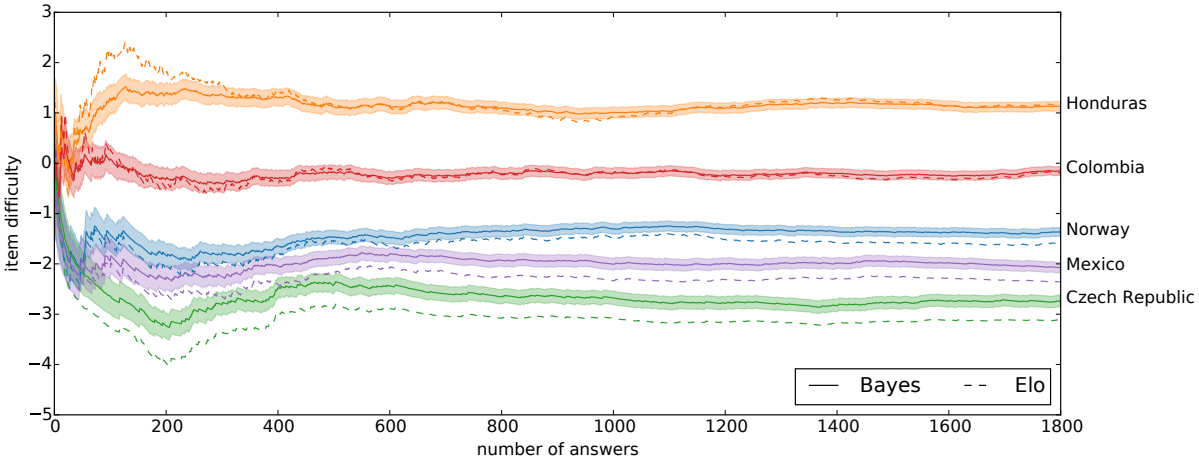

Figure 5: Difficulty of countries – the basic Elo model versus the Bayes model.

useful clusters of items. Such results can be used for improving the behavior of an adaptive educational system. For example, the system can let the user practice items from one concept and after reaching mastery move to the next one. Another possible use of concepts is for automatic construction of multiple-choice questions with good distractors (falling under the same concept).

We performed evaluation of the hierarchical model with different concepts. We used several approaches for specifying the concepts manually: based on type (e.g., countries, cities, rivers), location (e.g., Europe, Africa, Asia) and combination of the two approaches (e.g, European countries, European cities, African countries). Since we have most students' answers for European countries, we also considered a data set containing only answers on European countries. For this data set we used two sets of concepts. The first is the partition to Eastern, Western, Northwestern, Southern, Central and Southeastern Europe, the second concept set is obtained from the first one by union of Central, Western and Southern Europe (countries from these regions are mostly well-known by our Czech students) and union of Southeastern and Eastern Europe.

We compared these manually specified concepts with automatically corrected and entirely automatically constructed concepts (as described in Section 2.3; 'corrected' concepts are based on manually specified concepts and are revised based on the data). The quality of concepts was evaluated using prediction accuracy of the hierarchical model which uses these concepts. Table 2 shows the results expressed as RMSE improvement over the basic model. Note that the differences in RMSE are necessarily small, since the used models are very similar and differ only in the allocation of items to concepts. For the whole data set (1368 items) a larger number of concepts brings improvement of performance. The best results are achieved by manually specified concepts (combination of location and type of place), automatic correction does not lead to significantly different performance. For the smaller data set of European countries (39 items) a larger number of (both manual and automatically determined) concepts brings worse performance – a model with too small concepts suffers from a loss of information. In this case the best result is achieved by a correction of manually specified concepts. The analysis shows that the corrections make intuitive sense, most of them are shifts of well-known and easily recognizable countries as Russia or Iceland to block of well-known countries (union of Central, Western and Southern Europe).

Table 2: Comparison of manual, automatically corrected manual, and automatic concepts. Quality of concepts is expressed as RMSE improvement of the hierarchical model with these concepts over the basic model.

All items	number of concepts	RMSE improvement
manual – type	14	0.00132
corrected – type	14	0.00132
manual – location	22	0.00179
corrected – location	22	0.00167
manual – combination	**56**	**0.00235**
corrected – combination	56	0.00234
automatic	5	−0.00025
automatic	20	0.00039
automatic	50	0.00057
Europe		
manual	3	0.00003
corrected	**3**	**0.00011**
manual	6	−0.00015
corrected	6	0.00003
automatic	2	0.00007
automatic	3	0.00004
automatic	5	−0.00019

Models with multiple skills bring some additional information not just about the domain, but also about students. Correlation of concept skills with the global skill range from -0.1 to 0.5; the most correlated concepts are the ones with large number of answers like European countries (0.48) or

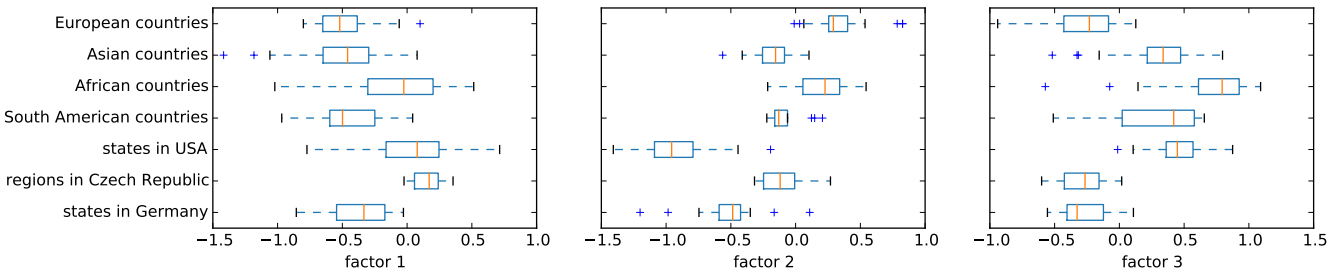

Figure 6: Boxplots of the item factor values from the Bayesian model (3 factors) grouped by some manually created concepts.

Asian countries (0.4), since answers on items in these concepts have also large influence on the global skill. Correlation between two clusters skills typically range from -0.1 to 0.1. These low correlation values suggest that concept skills hold interesting additional information about student knowledge.

Another view of relations between items is provided by the Bayesian model with multiplicative factors – this model does not provide division of items into disjoint sets, but rather determines for each item a strength of its relation to each factor (based on the data). Figure 6 illustrates how the learned factors relate to some manually specified concepts. Note that the results in Table 1 suggest that most of the improvement in predictive accuracy can be achieved by just these three automatically constructed factors. We can see that *factor 3* discriminates well between countries in Europe and Africa (Figure 7 provides a more detailed visualization). In the case of geography the division of items to concepts can be done in rather natural way and thus the potential application of such automatically determined division is limited and serves mainly as a verification of the method. For other domains (e.g., vocabulary learning) such natural division may not exist and this kind of model output can be very useful.

Also, note that *Factor 2* differentiates between states in USA and countries on other continents and *Factors 1* and *2* have different values for regions in Czech republic and states in Germany. This evidence supports an assumption that the model may be able to recognize students with varied background.

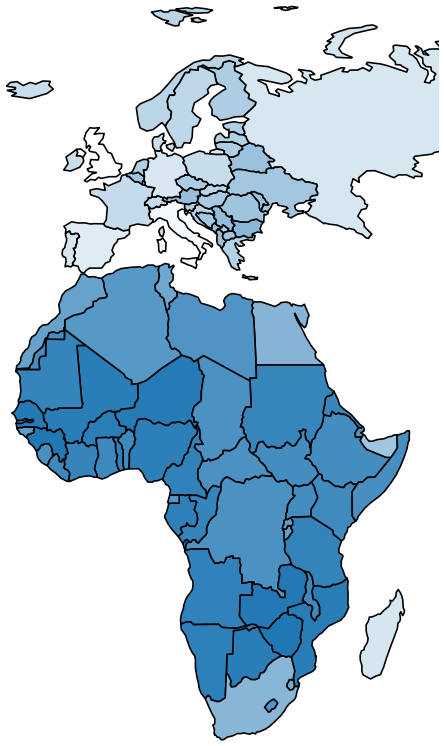

Figure 7: Visualization of the values of the third factor in the Bayesian model with multiple skills.

4. DISCUSSION

We have described and compared several student models of prior knowledge. The models were evaluated over extensive data from application for learning geography. The described models should be directly applicable to other online systems for learning facts, e.g., in areas like biology, human anatomy, or foreign language vocabulary. For application in domains which require deeper understanding (e.g., mathematics, physics) it may be necessary to develop extensions of described models (e.g., to capture prerequisite relations among concepts).

The results show that if we are concerned only with the accuracy of predictions, the basic Elo model is a reasonable choice. More complex models do improve predictions in statistically significant way, but the improvement is relatively small and evenly spread (i.e., individual predictions by different models are very similar).

The improvement in predictions by the hierarchical or networked models may be more pronounced in less homogeneous domains or with less homogeneous populations. Nevertheless, if the main aim of a student model is prediction of future answers (e.g., applied for selection of question), then the basic Elo model seems to be sufficient. Its performance is good and it is very simple to apply. Thus, we believe that it should be used more often both in implementations of educational software and in evaluations of student models.

The more complex models may still be useful, since improved accuracy is not the only purpose of student models. Described models have interpretable parameters – assignment of items to concepts and better quantification of uncertainty

of estimates of knowledge and difficulty. These parameters may be useful by themselves. We can use them to guide the adaptive behavior of educational systems, e.g., the choice of questions can be done in such a way that it respects the determined concepts or at the beginning of the session we can prefer items with low uncertainty (to have high confidence in choosing items with appropriate difficulty). The uncertainty parameter is useful for visualization of student knowledge in open learner models [2, 10]. Automatically determined concepts may also provide useful feedback to system developers, as they suggest potential improvements in user interface, and also to teachers for whom they offer insight into student's (mis)understanding of target domain. Given the small differences in predictive accuracy, future research into extensions of basic models should probably focus on these potential applications.

5. REFERENCES

[1] Tiffany Barnes. The q-matrix method: Mining student response data for knowledge. In *Educational Data Mining*, 2005.

[2] Susan Bull. Supporting learning with open learner models. In *Information and Communication Technologies in Education*, 2004.

[3] Cristina Conati, Abigail Gertner, and Kurt Vanlehn. Using bayesian networks to manage uncertainty in student modeling. *User modeling and user-adapted interaction*, 12(4):371–417, 2002.

[4] Albert T Corbett and John R Anderson. Knowledge tracing: Modeling the acquisition of procedural knowledge. *User modeling and user-adapted interaction*, 4(4):253–278, 1994.

[5] Michel C Desmarais, Behzad Beheshti, and Rhouma Naceur. Item to skills mapping: deriving a conjunctive q-matrix from data. In *Intelligent Tutoring Systems*, pages 454–463. Springer, 2012.

[6] Michel C Desmarais and Ryan SJ d Baker. A review of recent advances in learner and skill modeling in intelligent learning environments. *User Modeling and User-Adapted Interaction*, 22(1-2):9–38, 2012.

[7] Philipp Doebler, Mohsen Alavash, and Carsten Giessing. Adaptive experiments with a multivariate elo-type algorithm. *Behavior Research Methods*, pages 1–11, 2014.

[8] Jean-Paul Doignon and Jean-Claude Falmagne. *Knowledge spaces*. Springer, 1999.

[9] Arpad E Elo. *The rating of chessplayers, past and present*, volume 3. Batsford London, 1978.

[10] Carrie Demmans Epp, Susan Bull, and Matthew D Johnson. Visualising uncertainty for open learner model users. 2014. to appear.

[11] Mark E Glickman. Parameter estimation in large dynamic paired comparison experiments. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 48(3):377–394, 1999.

[12] Ralf Herbrich, Tom Minka, and Thore Graepel. Trueskill: A bayesian skill rating system. In *Advances in Neural Information Processing Systems*, pages 569–576, 2006.

[13] Tanja Käser, Severin Klingler, Alexander Gerhard Schwing, and Markus Gross. Beyond knowledge tracing: Modeling skill topologies with bayesian networks. In *Intelligent Tutoring Systems*, pages 188–198. Springer, 2014.

[14] S Klinkenberg, M Straatemeier, and HLJ Van der Maas. Computer adaptive practice of maths ability using a new item response model for on the fly ability and difficulty estimation. *Computers & Education*, 57(2):1813–1824, 2011.

[15] Yehuda Koren and Robert Bell. Advances in collaborative filtering. *Recommender Systems Handbook*, pages 145–186, 2011.

[16] Eva Millán, Tomasz Loboda, and Jose Luis Pérez-de-la Cruz. Bayesian networks for student model engineering. *Computers & Education*, 55(4):1663–1683, 2010.

[17] Thomas P Minka. *A family of algorithms for approximate Bayesian inference*. PhD thesis, Massachusetts Institute of Technology, 2001.

[18] Allan H Murphy. A new vector partition of the probability score. *Journal of Applied Meteorology*, 12(4):595–600, 1973.

[19] Juraj Nižnan, Radek Pelánek, and Jiří Řihák. Using problem solving times and expert opinion to detect skills. In *Educational Data Mining (EDM)*, pages 434–434, 2014.

[20] Juraj Nižnan, Radek Pelánek, and Jiří Řihák. Mapping problems to skills combining expert opinion and student data. In *Mathematical and Engineering Methods in Computer Science*, pages 113–124. Springer, 2014.

[21] Jan Papoušek and Radek Pelánek. Impact of adaptive educational system behaviour on student motivation. In *Artificial Intelligence in Education*, 2015.

[22] Jan Papoušek, Radek Pelánek, and Vít Stanislav. Adaptive practice of facts in domains with varied prior knowledge. In *Educational Data Mining (EDM)*, pages 6–13, 2014.

[23] Zachary A Pardos and Neil T Heffernan. Modeling individualization in a bayesian networks implementation of knowledge tracing. In *User Modeling, Adaptation, and Personalization*, pages 255–266. Springer, 2010.

[24] Philip I. Pavlik, Hao Cen, and Kenneth R. Koedinger. Performance factors analysis-a new alternative to knowledge tracing. In *Proc. of Artificial Intelligence in Education (AIED)*, volume 200 of *Frontiers in Artificial Intelligence and Applications*, pages 531–538. IOS Press, 2009.

[25] Radek Pelánek. Time decay functions and elo system in student modeling. In *Educational Data Mining (EDM)*, pages 21–27, 2014.

[26] Radek Pelánek. Metrics for evaluation of student models. *Journal of Educational Data Mining*, 2015. To appear.

[27] Ulrike Von Luxburg. A tutorial on spectral clustering. *Statistics and computing*, 17(4):395–416, 2007.